THE UNIVERSITY OF
MELBOURNE

Department of
Computer Science and Software Engineering

Concept of Operations

for

Managing the analysis of massive DNA sequence data

Version: 1.1
27 March 2009

This document aims to provide the audience a clear understanding of the problems faced by Peter MacCallum Cancer Centre, as well as the basic requirements of the desired piece of software, which the software development team attempts to achieve. Several other important factors in regards to the software and its stakeholders are also considered here.

Copyright notice

Department of Computer Science and Software Engineering
The University of Melbourne
Victoria
AUSTRALIA
3010

ICT Building
111 Barry Street
Carlton

Tel: +613 8344 1300
Fax: +613 9348 1184
https://www.csse.unimelb.edu.au/

Version: 1.1
27 March 2009.
https://www.cs.mu.oz.au/SE-projects/s340gd

Credits

This document was written by: Jack Low (wjlow), Lawrence Bang (lbang), Weng Hoe Ng (wengn), and Jinita Patel (jpatel).

Acknowledgments

The following people provided assistance with or were involved in the development of the document: Nick Chadwick and Jason Li.

CONTENTS

LIST OF FIGURES

ONE

Concept of Operations

1.1 Document Scope

This writing will begin by discussing the potential audience of the document, and why they would want to refer to it. It will also provide some background information on Peter MacCallum Cancer Centre and the problems that they are facing. The root causes of the problem will be explored and the possible solutions will be described. The aims of the software in development will be identified in relation to the problem at hand. Furthermore, the scope the software should cover will also be brought forward. This should give the audience an idea of the features of the completed software.

This writing relates to the problem at hand by bringing forward several important issues associated with the current system and the new system that is being developed.

The aims of the new system will be explained in relation to the current problem. In addition, a description of the causes of the problem will be explored to provide a better understanding of the aims of the new system.

1.2 Audience of Document

- The client (Jason Li):

This document will assist the client in confirming our understanding of the problem at hand. In the future, the client will also be able to use this document to refer back to what he has previously explained to the development team, and also of his expectations from that period.

- Supervisor of the development team:

This document will give the supervisor an indication of the development team's aims and understanding of the problem at hand. This will aid the supervisor in ensuring that the team is heading in the correct the direction.

- The development team:

The document will be used as a future reference to keep the project on track.

1.3 Background Information

One of the research activities conducted in Peter MacCallum Cancer Centre is to regularly compares short fragments of sampled DNA sequences to an existing library of known DNA sequences. This allows biologists to study genetic activities that have occurred in the sampled individual. BLAST is a program used to carry these processes out. Prior to applying BLAST, data is to be trimed as desired to remove junk data and also to be manipulated extract the regions of interest. One type of regions of interest is the unique identifier that was attached to the sequences during the experiment and the other type is the region comparable to the library of known DNA sequences. A piece of software has been requested to be built in order to simplify their job and to reduce operational costs.

Moreover, Peter MacCallum Cancer Centre is expecting more new data from external institutes to aid them in their research. This means that the amount of data they have to work with will eventually increase. This piece of software is meant to simplify their work in the future.

1.4 Root Causes of the Problem

- In order to match DNA sequences, R scripting is used to run many command lines and the BLAST processes need to be executed separately. This causes the sequencing process to take several hours to generate results, which are only shown after the whole process has been completed. The software in development will display results as it is being sequenced and matched, as well as aim to reduce the time by mapping several samples in one run.

- Most biologists are also unable to carry out this process of grouping and selecting DNA sequences on their own as they do not have the computer knowledge and the experience. Hence queries are passed to the client to be processed before being run through the BLAST. This causes one person to be responsible for all the queries that are sent by the biologists. The software in development will allow the biologists to sequence data on their own individual computers, as each individual biologist will have its own stand-alone computer (with the software) to work on.

- At the moment, only the client has the expertise and the experience to analyse the sequenced data. Many biologists, except for several bioinformaticians on-site, do not have the client's experience and so are unable to assist him in writing scripts for data. This causes delays in accessing the results and possibly their research. The software will enable biologists and bioinformaticians to easily adapt and learn the process of data without imposing the responsibility on one individual.

- The next generation sequencing takes AUD$4,000 to perform one experiment on one sample to obtain the 6 million records. To maximize the usage of one experiment, biologists would squeeze in multiple samples by attaching the unique identifiers to the samples before running the experiment. The software in development will allow the user to separate the data out by assessing the unique identifiers.

1.5 Software Aims

The primary aim of this piece of software is to reduce the wasteful consumption of funds and resources, which originates from the process of manually selecting, processing and grouping large DNA (nucleotide) sequences for the BLAST program. A graphical user interface (GUI) that uses BLAST is to be developed with additional functionality that will help to simplify the DNA matching process. Since BLAST takes several days to run, the program aims to take advantage of modern CPU's and enable multiple lines of DNA sequences to be processed by BLAST concurrently. Running this large amount (approximately 6 million lines) of data in parallel saves a lot of time. Furthermore, the program is intended to allow the separation of DNA sequences (identified by unique prefixes or suffixes) into several groups. This will allow different groups of DNA sequences to be treated differently, if necessary. For instance, sequences in group A could be BLASTed against library 1, while sequences in group B could be BLASTed against libraries 2 and 3.

The user should also be able to view the available results while BLAST is still running. This way it would be possible to view the number of matching sequences at any point during runtime, instead of having to wait until BLAST has finished processing all the data.

The current system employed by Peter MacCallum Cancer Centre involves a particular individual – Jason Li – manually selecting and processing data using self-written scripts. Any query from biologists would go through him in order to be processed and ran through the BLAST program. The piece of software in development is intended to solve this problem. It is to allow all biologists to be able to process any query on their own, in their own time.

1.6 Stakeholders

- In Peter MacCallum Cancer Centre

Biologists and bioinformaticians

They are the main users of the software, which is designed to help them out in their research.

Software Engineering and IT teams, e.g. programmers, designers, etc.

This software will be maintained by them in the future. Therefore, their work is expected to be affected by the software.

- In University of Melbourne

The institution.

The client approached the institution and requested the development of this software. Hence, the insitution holds some responsibility over the development process of this software.

Development team in charge of this software.

The development team's is responsible for meeting the requirements of the client.

Supervisor of the team.

The supervisor is to ensure the development team's successful completion of the software.

- Miscellaneous

Other cancer research institutes, biologists and bioinformaticians from all over the world.

Any advancements in cancer research caused by this software will assist them in their research in the future.

Cancer patients.

Advancements in cancer research caused by this software could lead to a change in their treatment process. Their DNA may also be used as input sequences for the software.

1.7 Boundaries and Constraints

Project constraints for the software:

a) Must be programmed in Java. b) Must be developed and ready to be delivered by the end of October(time constraint). c) Limited space available to store backups and process files. d) Download quota poses to be a limitation to the amount of data that can be received off the client.

Product constraints for the software:

a) Usability of the software will be constrained by the biologist due to their lack of knowledge of the software. b) Software has to be cross-platform, as biologists will be using PCs as well as Macs.

1.8 Document Overview

A user interface that wraps around the existing BLAST program has to be created. This should aid the user in selecting and manipulating query data before being passed through BLAST to be matched against a library of sequences, which are to be selected and manipulated by the user too. The program should show the user the number of input queries that match each sequence in the selected library.

A

Glossary

The following definitions, acronyms and abbreviations are used throughout this document:

1. **CSSE:** Department of Computer Science and Software Engineering
2. **GUI:** Graphical User Interface
3. **BLAST:** Basic Local Alignment Search Tool, or BLAST is a commonly used family of programs for matching DNA sequences.